利用词项语义共现和社团划分发现微博热点事件 *

李晓红，孔文文，马堉垠，马慧芳

(西北师范大学 计算机科学与工程学院，兰州 730070)

摘 要：针对传统词项之间语义关系抽取难以适用于微博，导致发现微博热点事件不敏感的问题，提出一种基于词项语义共现和社团划分的方法发现热点事件。首先利用热度定义对微博数据进行初次筛选，通过构建共现词项图来模拟词项间的语义相关性，并结合修改的 tf-idf 公式计算词项间的语义相关度；然后借助社区划分和模块度的概念对词项图进行划分，完成词项聚类，进而获得热点事件。实验结果表明，与同类方法相比，所提方法的准确率较高，发现的热点事件与实时事件基本保持一致，具有较好的热点识别效果。

关键词：热度；亲密度；语义相关性；热点事件；模块度

中图分类号：TP doi: 10.19734/j.issn.1001-3695.2018.09.0800

Title Microblog hot topic detection using lexical semantic co-occurrence and community partition

Li Xiaohong, Kong Wenwen, Ma Yuyin, Ma Huifang

(*College of Computer Science & Engineering*, *Northwest Normal University*, *Lanzhou* 730070, *China*)

Abstract: Due to difficulty to apply traditional method that extracts semantic relations between terms to microblog, which makes hot event detection not sensitive, this paper proposed a new method based on semantic co-occurrence of terms and community partition to find hot events. First, it utilized defined hotness to filter micro-blog data initially, then combined tf-idf formula with semantic relationships between items calculated by computing affinity score between two adjacent nodes on graph to harvest semantic relevancy between terms. Next, it introduced the idea of community partition to design the algorithm for word clustering, which made a series of microblog hot events finally obtained. Experimental results show the effectiveness of this method. Compared with kindred methods, this method has a higher accuracy, and hot event find is consistent with the real-time event basically, so this method can detect the microblog hot events effectively.

Key words: hot degree; affinity score; semantic relatedness; co-occurrence graph; modularity

0 引言

微博作为一种新兴的传播载体，已经成为民众表达舆情的重要窗口。它以简短快捷、内容丰富、用户“草根化”、传播速度快等特点，也成为了热点事件产生和讨论的重要场所。尤其是微博用户的关注、转发和评论等行为通常会助推微博事件的传播。随着微博信息泛滥成灾，大量有价值的数据被淹没，用户想找到自己感兴趣的话题变得力不从心。因此，如何从海量微博数据中挖掘出有价值的信息成为了计算机领域的研究热点。同时，微博热点事件发现作为网络舆情监控的重要分支，也受到了国内外学者的关注，具有重要的研究意义。

目前，针对微博热点事件发现的研究已有不少成果，主要可分为以下两类：

a）以文本为中心的方法[1]。先进行文本聚类，再在类中抽取出突发特征，从而识别突发事件。比如，陈羽中等人[3]提出 TCMLPA 聚类算法对微博的热点词语进行聚类，同时考虑聚类的时效，从而获得热点话题，并且提高了热点发现的精度。文献[3, 4]分别致力于不同的聚类算法，如 K_SC 聚类算法、SEPPM 模型，对网络热点事件进行发现和提取，取得了一定的成效。但是一方面，由于微博内容的简短，严重的数据稀疏问题；另一方面，微博含有很多噪声数据，聚类后在识别突发词的效率就会比较低下。

b）以突发特征为中心的方法。先抽取突发特征并对其进行分组，然后使用突发特征组进行突发事件的识别。Fu 等人[5]基于语言和主题模型，通过相邻时间间隔之间的情绪分布语言模型的差异来发现微博热门话题，Yang 等人[6]使用基于时间窗的分析方法来检测突发特征，然后使用相似度传播 AP 算法对突发特征进行聚类。类似地，贺敏等人[7]对关键词定义加权公式，并引入滑动窗口，以实时监控热点事件的发生。刘业政等人[8]提出利用单个话题构建表征话题属性的热度曲线，然后对热度曲线进行分类建模，最后在分类模型上使用加权投票规则来预测新话题是否会发展成为热门话题。上述几种方法在检测突发事件时只是理论上提高了事件发现的性能，但在实际应用中并不能得到很好的话题发现效果，其最根本原因是事件发现过程中，话题会随时间变化产生话题漂移现象。

为了提高微博热点事件检测的准确性并降低复杂度，本文提出了一种基于特征词语义相关性和社团结构的微博热点事件发现算法(using lexical semantic co-occurrence and community structure to find microblog hot event， LSCaCS)。具体地，通过构建无向带权词项图获取词项之间潜在的语义关系，并计算语义强度，然后利用社团发现和模块度思想在图上完成词的聚类，实现热点词与热点事件的对应。算法流

收稿日期：2018-09-30；**修回日期：**2018-12-02 **基金项目：**国家自然科学基金资助项目（61862058，61762078）；甘肃省青年基金资助项目（1606RJYA269）；西北师范大学青年教师科研能力提升计划项目（NWNU-LKQN-14-5，NWNU-LKQN-16-20）

作者简介：李晓红（1978-），女，兰州市人，讲师，硕士，主要研究方向为数据挖掘、机器学习（xiaohongli@nwnu.edu.cn）；孔文文（1998-），女，本科生；马堉垠（1997-），女，硕士研究生；马慧芳（1981-），女，副教授，博士，主要研究方向为机器学习、数据挖掘.

程如图 1 所示。

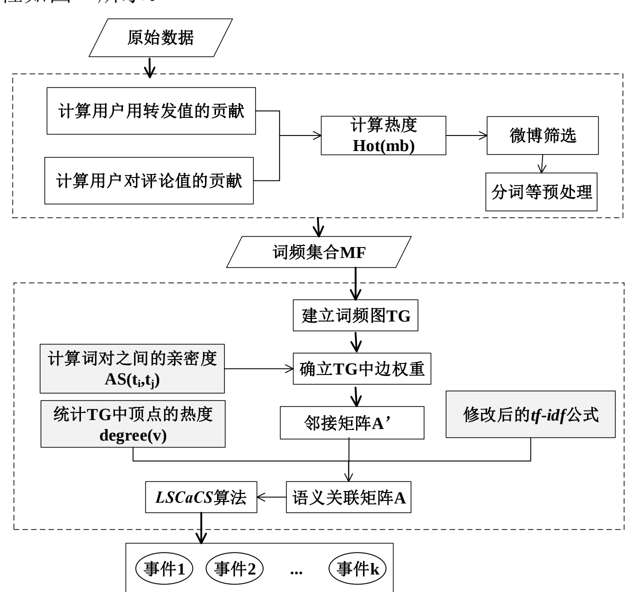


图 1　微博热点事件发现算法流程

Fig. 1　Flow chart of micro-blog hot event detection algorithm

1　相关知识

首先给出本文用到的符号定义，如表 1 所示。

表 1　符号含义表

Table 1　Notations

符号	定义
$MB=\{mb_1,mb_2,\ldots,mb_N\}$	微博的数据集
$MT=\{t_1,t_2,\ldots,t_M\}$	特征词集
mb_i	第 i 条微博
N	微博数据集大小
M	特征词数目，图 G 的顶点数目
$AS(t_i, t_j,mb)$	t_i 与 t_j 在微博 mb 中的亲密度

1.1　微博的热度

微博对热点事件的敏感性使其在一定程度上可以反映热点事件。一般来说，关注度高的微博，其转发数和评论数会逐渐上升，且在较短时间内传播。因此，需要一个指标来度量微博被关注的程度。

假设用户 u_i 发布了一条微博 mb，被用户 u_j 转发，则用户 u_j 对微博转发值的贡献记为 $cret(mb, u_j)$。

$$cret(mb,u_j)=\begin{cases}1 & \text{否则}\\ 1-f(u_i,u_j) & \text{若}u_j\text{是}u_i\text{粉丝}\end{cases} \tag{1}$$

同理，用户 u_j 对微博评论值的贡献用 $ccom(mb, u_j)$表示如下：

$$ccom(mb,u_j)=\begin{cases}1 & \text{否则}\\ 1-f(u_i,u_j) & \text{若}u_j\text{是}u_i\text{'粉丝}\end{cases} \tag{2}$$

其中：$f(u_i,u_j)=\dfrac{count(u_j)}{count(u_i)}$ 定义为用户 u_j 对用户 u_i 的关注度；$count(u_i)$表示用户 u_i 所关注用户的数目。基于式(1)和(2)，给出热度的定义。

定义 1　热度。　热度指在单位时间内，所有用户对该微博的转发值贡献 $cret(mb, u_j)$与评论值贡献 $ccom(mb, u_j)$加权和的平均值。

$$Hot(mb)=\frac{\lambda\sum_{j=1}^{l}cret(mb,u_j)+(1-\lambda)\sum_{j=1}^{h}ccom(mb,u_j)}{(l+h)\Delta t} \tag{3}$$

其中：λ 为调节参数，且 $0<\lambda<1$，l 为微博 mb 的转发次数；h 为评论次数。通过热度的定义可以初步判断微博所描述的事件成为热点事件的可能性，且热度与微博内容无关。

1.2　词项间的亲密度

定义 2　d 度距离。　设函数 $distance(t_i, t_j, mb)$可计算微博 mb 中特征词 t_i 与 t_j 之间所包含的词项个数。若式(4)成立，则定义微博 mb 中特征词 t_i 与 t_j 之间的距离为 d 度距离。

$$d=distance(t_i,t_j,mb) \tag{4}$$

例如，微博 mb1=“腾讯公司董事会主席兼 CEO 马化腾”分词后所得词项集合为:{腾讯，公司，董事会，主席，CEO，马化腾}，则 *distance*(腾讯，马化腾, *mb*1)=4，表示在 *mb*1 中“腾讯”和“马化腾”之间为 4 度距离；*distance*(腾讯，公司, *mb*1)=0，表示 “腾讯”和“公司”之间为 0 度距离。

定义 3　亲密度。　给定微博 mb，若词项 t_i 与 t_j 之间为 d 度距离，则词项 t_i 与 t_j 在 d 度距离上的亲密度定义为式(5)中 $AS(t_i, t_j)$的值。

$$AS(t_i,t_j)=n_d\times e^{-d} \tag{5}$$

其中：n_d 表示在数据集中特征词 t_i 与 t_j 之间距离为 d 度距离的微博数。

词的亲密度 $AS(t_i,t_j)$ 意味着：a) 如果两个词项共现的距离不同，则它们亲密的程度会有所不同；b) 如果两个词经常共同出现，则这两个词在意义上是相互关联的，$AS(t_i, t_j)$越高，关系越紧密。与传统的共现强度计算比较，该方法更为合理[9]。

1.3　模块度

许多大规模复杂网络是由若干个“社区”或“组”构成的。一个相对好的划分是每个社团内部节点间的连接非常紧密，社团之间的连接相对比较稀疏。而模块度[10]就是衡量一个社区划分好坏的常用指标，计算公式如下：

$$Q=\frac{1}{2m}\sum_{i,j}[P_{ij}-\frac{k_ik_j}{2m}]\delta(C_i,C_j) \tag{6}$$

其中：$m=\frac{1}{2}\sum_{i,j}P_{ij}$ 表示图中所有边上的权重之和；P_{ij} 表示顶点 i 和顶点 j 之间边上的权重；$k_i=\sum_j P_{ij}$ 表示所有与节点 i 相连的边的权重之和；C_i 表示节点 i 所属的社区。

$$\delta(C_i,C_j)=\begin{cases}0 & i=j\\ 1 & otherwise\end{cases} \tag{7}$$

Q 值在 0~1 之间。Q 值越大，图划分的社区结构准确度越高，Q 最大时说明图划分较理想。

2　发现热点事件

2.1　构建词项图并获取语义相关度

将 MT 中的词映射为图中的顶点，词项之间的共现关系用无向带权图 $TG=(V, E)$ 来表示，则顶点集合为 $V=\{v_1,v_2,\ldots,v_M\}$，其中顶点 v_i 为特征词项（注：本文余下部分使用 v_i 表示词 t_i）。如果两个词 v_i、v_j 来自同一微博，则将 v_i、v_j 之间相连构成一条边(v_i, v_j)。设图 TG 的邻接矩阵表示为 A'，A' 中的元素记为 w'_{ij}。

首先，计算 A' 中元素 w'_{ij} 的值。w'_{ij} 是边(v_i, v_j)上的权值，表示顶点 v_i，v_j 在微博数据集中的语义相关度，其值可通过它们在数据集中的亲密度之和计算得到[12]。考虑到距离过大时，特征词之间的共现对它们的亲密度没有意义，故本文实验取 $0\leq d\leq 6$。

$$w'_{ij}=\begin{cases}\sum AS(v_i,v_j) & (v_i,v_j)\in E\\ 0 & otherwise\end{cases} \tag{8}$$

$w'_{ij}=w'_{ji}$，A' 为对称矩阵。接下来，对顶点之间的语义

关联关系进行归一化和非对称化处理。通常，如果两个词在数据集中越是频繁且近距离的共同出现，则它们之间的语义关联性就越高；同时，那些几乎与所有特征词都具有亲密关系的词项是没有意义的，即这样的亲密关系在语义相关度模型中并不重要，必须进行惩罚。因此，综合考虑词项亲密度对微博的重要性和其在数据集中的普遍性，最大程度地挖掘出亲密度对微博热点事件检测的语义贡献。同时结合文档逆文档频率(*tf-idf*)所表示的含义，对其计算公式进行适当更改[11]，推得式(7)。

$$w_{ij}=\frac{w'_{ij}}{\sum_j w'_{ij}}\bullet\log\frac{M}{degree(v_i)} \tag{9}$$

其中：$degree(v_i)$表示顶点 v_i 的度，也就是与词项 v_i 具有亲密关系的总词数；$\log\frac{M}{degree(v_i)}$ 用来惩罚几乎与所有词项都亲密的特征词。将式(7)应用于 A' 中元素的值[11]，最后得到非对称矩阵 A。

$$A=\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1M} \\ w_{21} & w_{22} & \cdots & w_{2M} \\ \cdots & \cdots & \cdots & \cdots \\ w_{M1} & w_{M2} & \cdots & w_{MM} \end{bmatrix}$$

其中：$w_{ij}\neq w_{ji}$，且 $0\le w_{ij}\le 1$，w_{ij} 越大，该词对的主题预测能力就越强，特别地，$w_{ij}=0$，可将图 *TG* 中对应词对之间原有的边删除，以此达到简化图结构、降低运算复杂度的目的。本文后面的内容将基于图 *TG* 展开。

2.2 热点事件发现算法描述

热点事件发现通常使用的方法大部分以文档为聚类对象，聚类结果为文档簇。而本文利用社区发现思路，以词项图作为划分对象，结合词项间的语义关联关系，简化聚类过程，提出了一种基于特征词语义相关性和社团结构的微博热点事件发现算法 *LSCaCS*，从而达到发现热点事件的目的。

首先，初始化社区结构，即将图 *G* 中每一个顶点均看做为一个独立的社区，先在矩阵 *A* 中查找最大值，假设 $max(A)=w_{ij}$，将 w_{ij} 对应的两个顶点划分在同一个社区；然后，以这对节点为种子扩展社区，每扩展一个节点计算一次模块度增量ΔQ，若$\Delta Q>0$，则扩展成功。重复这个过程，直到 *nodestack* 为空为止。栈 *nodestack* 用来保存待扩展的社区节点，用 *processed* 保存已发现的社区顶点。具体的算法 *LSCaCS* 步骤如下：

输入：语义关系矩阵 A，参数 β。

输出：一个社区。

1：初始化 nodestack=Φ, processed=Φ；

2：查找矩阵 A 中元素的最大值，若 $\max(A)=w_{ij}$，则执行入栈操作：push(nodestack,(v_i, i))，push(nodestack,(v_j, j))；

3：置 $w_{ij}=0$；

4：循环，重复执行以下操作，直到 *nodestack* 为空:

4.1 出栈：pop(nodestack, (v_r, r));

4.2 processed= processed ∪ v_k;

4.3 循环：对 A 中第 r 行的元素依次执行下列操作，

if $w_{rk}>\beta$ 且 $v_k\notin processed$，则执行：

4.3.1 计算 $v_k \cup processed$ 后所构成社区的ΔQ;

4.3.2 if $\Delta Q>0$，则 push(nodestack, (v_k, r));

5：return processed. //结束

本算法结束后得到一个社区，即对应一个事件簇，并取 *top-K* 的权重对应的顶点来描述热点事件。若要获取所有的热点事件，则需更新整个网络。重复上述操作，直到所有的热点事件被发现。

3 实验结果及分析

3.1 实验数据

人工采集了新浪微博从 2017 年 1 月~6 月发表的微博作为实验数据。为保证与真实话题最大程度上的一致性，采样时人工加入了适量的噪声数据，构造了一个共包含 3 225 条微博、8 类热点事件的有噪声的微博数据集，其中描述事件的微博 2 541 条，噪声数据 684 条。对其进行数据清洗、分词、去停用词等预处理操作，并根据词项之间关系的紧密性进行了孤立词筛选，最终保留了 28 600 个词项。实验数据集如表 2 所示。

表 2 实验数据集

Table 2 Experimental data sets

类别	微博数	平均评论数	平均转发数	类别	微博数	平均评论数	平均转发数
事件 1	387	1566	526	事件 5	366	558	212
事件 2	402	3121	957	事件 6	390	2026	336
事件 3	406	1314	6862	事件 7	421	951	489
事件 4	440	8175	7448	事件 8	413	1292	1025

3.2 评价指标

本文引进 *NMI*[12]和 *ARI*[13]两个评价指标对实验结果进行综合评价。并设真实类别为 $C=\{c_1,\ldots c_s,\ldots\}$，聚类结果为 $\Omega=\{\omega_1, \ldots\omega_k, \ldots\}$。

NMI 的取值为[0, 1]，*NMI* 值越大，表示事件发现的结果越接近真实情况。则 *NMI* 的定义如下，其中：$p(\omega_k,c_j)$ 是联合概率；$I(\Omega,C)$ 是互信息。

$$I(\Omega,C)=\sum_k\sum_j p(\omega_k,c_j)\log\frac{p(\omega_k,c_j)}{p(\omega_k)p(c_j)} \tag{10}$$

$$NMI(\Omega,C)=-2*\frac{I(\Omega,C)}{\sum_k p(\omega_k)\log p(\omega_k)+\sum_j p(c_j)\log p(c_j)} \tag{11}$$

ARI 也用来评价聚类的效果，取值为[−1，1]，衡量的是两个数据分布的吻合程度，值越大，意味着聚类结果与真实情况越吻合。定义如式(13)所示。

$$ARI=\frac{RI-E(RI)}{\max(RI)-E(RI)} \tag{12}$$

$$RI=\frac{a+b}{C_n^2} \tag{13}$$

其中：a 表示在 C 与 Ω 中都是同类别的元素对数；b 表示在 C 与 Ω 中都是不同类别的元素对数。

3.3 实验结果与分析

通过将发现的事件结果与真实发生的网络事件相比较来评价方法的性能。本文设计了三组实验来验证热点事件发现算法的有效性。实验 1 调整微博热度和热点事件发现算法中重要的参数值，以观察对热点事件结果的影响；实验 2 利用数据集抽取了热点话题，并与真实热点事件进行了比较；实验 3 对本文方法与已有同类方法的热点事件检测结果进行了对比。

实验 1 研究参数 λ 和 β 取不同值对热点事件中主题词提取结果的影响。参数 λ 权衡 $cret(mb, u_j)$和 $ccom(mb, u_j)$这两个因素对微博热度的影响；β 考虑特征词之间关系的紧密程度(语义相关度)对热点主题词提取贡献的大小。β 从 0.01 到 0.08 变化，λ 选取了三个值，分别是 0.48、0.5 和 0.52。实验结果如图 2、3 所示。

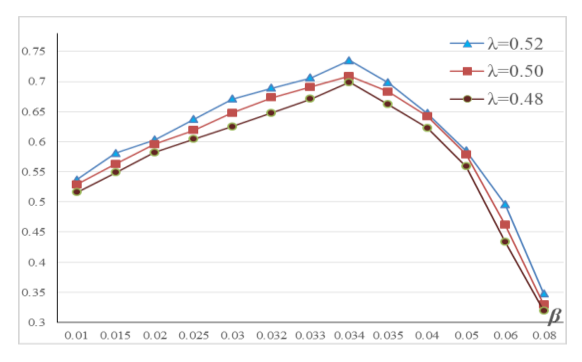


图 2　参数对 NMI 的影响

Fig. 2　Influence of varied parameters on *NMI*

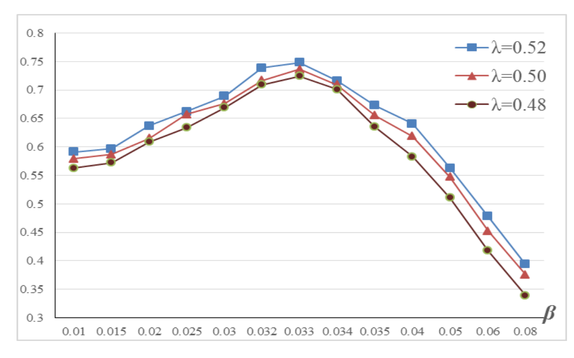


图 3　参数对 ARI 的影响

Fig. 3　Influence of varied parameters on *ARI*

从图 2 和 3 均可以看出，一方面，λ 取值不同时对应的 *NMI* 和 *ARI* 也不相同，就意味着热点事件的检测中，微博转发值和评论值对其贡献不同，本实验结果发现转发值的贡献更大。另一方面，当 $\beta\leq0.034$(图 3 中 $\beta\leq0.33$)时，*NMI* 和 *ARI* 的性能曲线均呈上升趋势；当 $\beta=0.034$($\beta=0.33$)时，*NMI* 和 *ARI* 达到最高值。但是随着 β 值的持续增加，性能曲线趋于下降。特别地，当 β 超过 0.05(0.04)后，下降速度变得更快，说明词项间关系的亲密程度对热点事件主题词的准确度有较大的影响。

表 3　本文方法检测出的热点事件

Table 3　Comparison between real events and hot events of this method detects

真实热点话题	本文检测出的热点词汇
百度不再是联网公司	百度　互联网　公司　*AI*　人工智能　李彦宏
肉松饼里的肉松是棉花	肉松饼　棉花　泡水　老婆饼　燃烧
黄渤完全可以去说相声	幽默　尴尬　说相声　黄渤　岳云鹏　情商
《欢乐颂》安迪两任男友	欢乐颂　安迪　男友　喜欢　爱　差别　小包总
尔康制药销售藏惊人秘密	尔康　销售　秘密　经销商　紫薇　制药销售
儿子染上毒品和赌博	毒瘾　儿子　痛心　父母　卖房　还债
柯洁高度评价 *AlphaGo*	围棋　柯洁　*AlphaGo*　评价　输棋
女子酒店遇袭事件	完美　女子　酒店　电梯　遇袭　真相

实验 2　基于实验 1 的结果，本实验设定参数的值为 $\lambda=0.475$，$\beta=0.03$，构建基于微博数据的词项之间的语义关系矩阵，通过 LSCaCS 算法对大量微博所涉及的热点事件进行提取。

实验结果如表 3 所示。实验选取所得事件簇中足以描述事件主题的词项来表征发现的热点事件，并与权威机构发布的热点事件进行了对比。结果显示，本文方法发现的热点事件与真实网络上的热点事件基本吻合，说明本文所提出的发现热点话题的算法是有效的。

实验 3　选择文献[14]的离散粒子群优化(*DPSO*)算法和文献[15]提出的组合模型方法(*MCF*)与本文的 *LSCaCS* 方法在数据集上进行了实验和结果对比分析。其中，*DPSO* 算法通过对词语互信息及内外关联词信息的挖掘，利用离散粒子群优化算法从寻优角度发现微博热点话题。*MCF* 方法提出使用主题模型提取出微博主题，引入词激活力模型计算词之间的词激活力，利用词激活力矩阵生成热点事件。实验对比结果如图 4 所示。

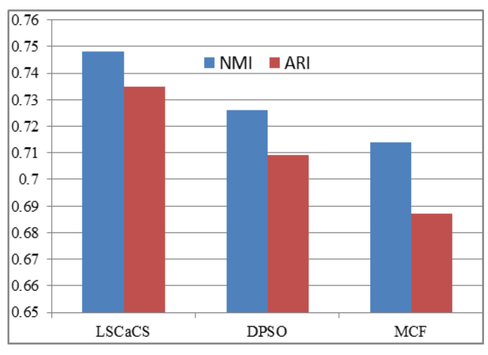


图 4　三种方法实验结果对比

Fig. 4　Comparison among experimental results of three methods

由图 4 可以看出，与其他两种方法相比，本文方法的 *NMI* 和 *ARI* 略高。首先，本文所提的方法充分挖掘了词项之间表层和隐含的语义关系，不但考虑词项之间的共现，而且还考虑了不同亲密关系的共现，使得微博语义表示更清楚更仔细；其次，基于原始数据集在构建推导词项语义关系矩阵的过程中，将一些不重要的数据进行了删除，故干扰较少。基于以上原因，使得本文方法得到的结果较好。因微博有内容短、表述不规范、噪声多等缺陷，导致其他两种方法选取的主题词数量不足、质量不高，最终导致事件发现的结果不佳。

4　结束语

本文提出了一种基于特征词语义相关性和社团结构的微博热点事件发现算法，主要设计思路是通过构建无向带权词项图获取词项之间显示的和隐含的语义关系，计算语义强度，并构建语义关联关系矩阵；同时，引入社区划分的思想，利用 *LSCaCS* 算法对词项进行聚类，从而获得热点事件的集合。实验结果表明，发现的热点话题与实时事件保持一致，具有较好的热点识别效果。今后可以围绕降低特征词集中离群点的数量、随机游走模型指标的初始化以及社团划分收敛条件的判定标准进行相关研究，甚至可以尝试引入专业领域词典或词汇本体，进而提升热点事件发现的准确性。

参考文献：

[1] Qi Xiang, Huang Yu, Chen Ziyan, *et al.* Burst-LDA: a new topic model for detecting bursty topics from stream text [J]. Journal of Electronics, 2014, 31 (6): 565-575.

[2] Xie Wei, Zhu Feida, Jiang Jing, *et al.* Topic sketch: real-time bursty topic detection from twitter [J]. IEEE Trans on Knowledge and Data Engineering, 2016, 28 (8): 2216-2229.

[3] 陈羽中，方明月，郭文忠. 面向微博热点话题发现的多标签传播聚类方法研究 [J]. 模式识别与人工智能, 2015, 28 (1): 1-10. (Chen Yuzhong, Fang Mingyue, Guo Wenzhong. Research on multi-label propagation clustering method for microblog hot topic detection [J]. Pattern Recognition and Artificial Intelligence. 2015, 28 (1): 1-10.)

[4] Spyrou E, Mylonas P. Analyzing flickr metadata to extract location-based information and semantically organize its photo content

[J]. Neurocomputing, 2016, 172: 114-133.)
[5] Fu Xianghua, Li Jianqiang, Yang Kun, *et al.* Dynamic online HDP model for discovering evolutionary topics from Chinese social texts [J]. Neurocomputing, 2016, 171: 412-424.
[6] Yang Liang, Lin Hongfei, Lin Yuan, *et al.* Detection and extraction of hot topics on chinese microblogs [J]. Cognitive Computation, 2016, 8 (4): 577-586.
[7] 贺敏，徐杰，杜攀,等．基于时间序列分析的微博突发话题检测方法 [J]．通信学报, 2016, 37 (3): 48-54. (He Min, Xu Jie, Du Pan, *et al.* Bursty topic detection method for Microblog based on time series analysis [J]. Journal on Communications, 2016 ，37(3): 48-54.)
[8] 刘业政，杜亚楠，姜元春，等．基于热度曲线分类建模的微博热门话题预测 [J]．模式识别与人工智能, 2015, 28 (1): 27-34. (Liu Yezheng, Du Yanan, Jiang Yuanchun, *et al.* Trend prediction for Microblog based on classification modeling of heat curves [J]. Pattern Recognition and Artificial Intelligence, 2015, 28 (1): 27-34.)
[9] Pan Jiayu, Yang H J, Duygulu P. *et al.* Automatic multimedia cross-modal correlation discovery [C]// Proc of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]:ACM Press, 2004: 653-658.
[10] Newman M E J, Girvan M. Finding and evaluating community structure in networks [J]. Physical review E, 2004, 69 (2): 026113.
[11] Wen Hua, Wang Zhongyuan, Wang Haixun, *et al.* Short text understanding through lexical-semantic analysis [C]// Proc of the 31st IEEE International Conference on Data Engineering. [S.l.]:IEEE Press, 2015: 495-506.
[12] Rand W M. Objective criteria for the evaluation of clustering methods [J]. Journal of the American Statistical association, 1971, 66 (336): 846-850.
[13] Fahad A, Alshatri N, Tari Z, *et al.* A survey of clustering algorithms for big data: taxonomy and empirical analysis [J]. IEEE Trans on Emerging Topics in Computing, 2014, 2 (3): 267-279.
[14] 马慧芳，吉余岗，李晓红，等．基于离散粒子群优化的微博热点话题发现算法 [J]．计算机工程, 2016, 42 (3): 208-213. (Ma Huifang, Ji Yugang, Li Xiaohong, *et al.* Hot topic discovering algorithm for microblog based on discrete particle swarm optimization [J]. Computer Engineering, 2016, 42 (3): 208-213.)
[15] 戴天，吴渝，雷大江．利用组合模型生成微博热点话题事件摘要 [J]. 计算机应用研究, 2016, 33 (7): 2026-2029. (Dai Tian, Wu Yu, Lei Dajiang. Hot topic summarization on Microblog generated by model combination [J]. Application Research of Computers. 2016, 33 (7): 2026-2029.)